Simen Owen Academic Proceedings Series

3rd International Conference on Electronics, Engineering, Computer Science and Applied Development (EESD 2026)

Article

Retrieval-Augmented Vision-Language-Action Policies for Cross-Embodiment Generalization in Robotic Manipulation

Jiacheng Geng[1,*]

[1] Zhejiang University-University of Illinois Urbana-Champaign Institute, Haining, China
* Correspondence: Jiacheng Geng, Zhejiang University-University of Illinois Urbana-Champaign Institute, Haining, China

Abstract: Robotic manipulation is gradually evolving from performing tasks in fixed scenarios to the flexible deployment of robots, tools, and complex environments. The Vision-Language-Action (VLA) paradigm enables robots to generate executable actions by integrating visual perception with natural language instructions. However, when robots differ in kinematic structures, end-effector configurations, observation perspectives, or action scales, the transferability of learned policies remains severely limited. Existing VLA methods and retrieval-based strategies typically reuse historical task experiences without considering whether those experiences are physically compatible with the current robot embodiment. To address this challenge, this study proposes an enhanced retrieval-augmented VLA framework that explicitly incorporates robot body information into the experience retrieval process. The proposed method improves the applicability of historical trajectories through compatibility screening and target robot action-space mapping. Specifically, the framework retrieves similar operation demonstrations by jointly integrating visual, linguistic, and embodiment features, then selects the most effective experiences based on the physical characteristics of the target robot to generate corresponding actions. Five experiments were conducted on a subset of the publicly available Open X-Embodiment dataset. Results demonstrate that, compared with the strongest baseline, the proposed method achieves notable improvements in action prediction accuracy, cross-embodiment transfer capability, and vision-language retrieval compatibility. These findings indicate that retrieval methods that account for a robot's own physical characteristics can effectively enhance action prediction accuracy, task transferability, and result interpretability in cross-embodiment robotic operations.

Keywords: robotic manipulation; vision-language-action policy; retrieval-augmented learning; cross-embodiment generalization; embodiment-aware retrieval

Received: 29 May 2026
Revised: 15 July 2026
Accepted: 26 July 2026
Published: 01 August 2026

1. Introduction

Robotic operation is progressively evolving from executing tasks in static, controlled environments to supporting adaptable applications across diverse robotic platforms, end-effectors, and task configurations. Vision-language-action frameworks enable the mapping of visual inputs and natural language directives into executable control signals, thereby exhibiting substantial promise for real-world deployment. A persistent challenge in practical robotic systems remains the limited transferability of learned strategies across different robot platforms [1]. Variations among robots—including degrees of freedom, mounting configurations, camera viewpoints, control loop frequencies, and action scaling—pose significant barriers to cross-platform generalization in robot learning, particularly when acquiring large-scale task-specific training data for newly deployed robots entails considerable time and resource investment.

Recent advances in large-scale robot learning, imitation-based approaches, and vision-language-action-driven operation methods have yielded measurable improvements in task-level generalization. However, several critical limitations persist. First, many vision-language-action frameworks rely predominantly on instantaneous visual input and linguistic commands during execution, with underutilization of previously acquired operational knowledge. Second, existing retrieval-based techniques typically overlook whether retrieved historical trajectories align with the physical embodiment and actuation capabilities of the target robot. Consequently, even trajectories that appear semantically or visually similar may prove infeasible due to structural or kinematic incompatibilities [2]. Third, current cross-platform adaptation strategies often depend on model fine-tuning or standardized action representations, yet lack transparent mechanisms to identify which prior experiences are genuinely beneficial for the current robot—and why—thereby limiting interpretability and reliability.

To address these challenges, this study introduces a retrieval-enhanced vision-language-action framework tailored for cross-embodiment robotic operation. The central premise is to leverage stored operational experience during task execution, enabling heterogeneous robots to share and repurpose historical trajectories effectively. The methodology comprises three integrated components: first, retrieval incorporates visual observations, language instructions, and robot-specific embodied features—including joint limits, end-effector geometry, and sensor modalities—to select historically validated trajectories most compatible with the target platform; second, the retrieved trajectories serve as contextual priors, jointly encoded with current sensory and linguistic inputs to refine action prediction within the vision-language-action model; third, the predicted actions undergo kinematic and dynamic adaptation to conform to the motion constraints and control interface of the target robot, ensuring safe and executable behavior across platforms. Empirical validation employs comparative analysis of alternative retrieval paradigms alongside systematic ablation studies to assess performance gains [2].

The overall workflow follows a streamlined pipeline: multimodal encoders process visual inputs, language instructions, and robot embodiment descriptors; an experience database is queried to retrieve the most relevant prior operational instances aligned with the current task context; the selected historical trajectories are fused with real-time observation data and fed into the vision-language-action model to generate action sequences; finally, output actions are transformed to match the target robot's action space and translated into low-level control commands suitable for execution [3].

This work contributes a novel paradigm for enhancing cross-embodiment robot learning. Theoretically, it advances understanding of how robot-specific haptic and kinematic attributes inform experience retrieval—offering methodological insights for future research in embodied AI. Practically, the approach reduces dependence on extensive retraining data when integrating new robotic hardware, thereby improving strategy portability across platforms. By prioritizing retrieval of trajectories demonstrably compatible with the target robot's physical configuration, the method mitigates execution risks arising from structural mismatches, yields actions better aligned with the robot's intrinsic dynamics, and strengthens the robustness and fidelity of autonomous decision-making [3].

2. Related Works

2.1. Vision-Language-Action Policies for Robotic Manipulation

The vision-language-action strategy has become an important research direction in robotic operation. Its core idea is to integrate visual perception, language understanding, and action generation into a unified framework [4]. Compared with traditional methods that rely on predefined states, artificial rules, or task-specific controllers, this strategy can jointly process image observations and natural language instructions to generate robot actions. Therefore, it holds promising potential for tasks such as grasping, placing, relocating objects, and basic tool interaction.

Current vision-language-action approaches remain highly dependent on the distribution of training data [5]. When the target robot differs from those used during training—such as in manipulator configuration, end-effector design, camera placement, control frequency, or action scaling—the generated actions may satisfy task objectives but fail to execute directly on the target platform. Enhancing cross-robot generalization capability thus remains a key challenge in ongoing research.

2.2. Cross-Embodiment Robot Generalization

Cross-embodiment robot learning seeks to transfer operational knowledge across diverse robotic platforms, thereby decreasing the dependence of newly deployed robots on extensive task-specific data [6]. Current approaches primarily accomplish this by sharing feature representations, standardizing action spaces, or incorporating contextual cues to reduce inter-robot discrepancies and improve the portability of control policies.

Although existing methods acknowledge the role of embodiment, embodied information is often underutilized—frequently treated merely as a static input condition [7]. In reality, distinct robots may execute the same task through divergent kinematic and dynamic patterns, making it unreliable to assume direct trajectory transferability based solely on shared representations. Consequently, many cross-embodiment techniques still necessitate model adaptation for each new platform, increasing deployment overhead and diminishing model transparency.

2.3. Retrieval-Augmented Robot Learning

Enhanced learning offers a novel paradigm for robot operations. Rather than relying exclusively on model parameters and immediate sensory inputs to generate actions, these methods leverage historical trajectories or task demonstrations stored in an external experience database to inform robotic decision-making [7]. In operational tasks, retrieved examples encode information such as object configurations and procedural sequences, thereby supplying actionable guidance for action generation. Furthermore, because the model's decisions can be explicitly grounded in specific prior instances, retrieval-enhanced approaches typically exhibit improved interpretability.

However, most existing retrieval strategies select historical experiences based on similarity at linguistic, visual, or task levels, with limited consideration of compatibility with the target robot's physical configuration or kinematic capabilities [8]. Consequently, retrieved examples—while semantically relevant to the current task—may not be directly executable due to disparities in robotic morphology or actuation characteristics. This limitation is especially pronounced in heterogeneous robot settings, where identical task specifications do not guarantee transferable execution strategies across distinct robotic platforms.

2.4. Research Gap and Position of This Study

Overall, the vision-language-action strategy, cross-situational learning, and retrieval-enhanced learning have each advanced robot operations in task understanding, strategy transfer, and experience utilization, respectively. However, integrated application of these three dimensions remains limited. Most existing studies address only one dimension in isolation and seldom examine the interplay among embodied information, experience retrieval, and action generation within a unified framework [2].

This paper introduces a vision-language-action framework enhanced by retrieval mechanisms, unifying task understanding, historical experience utilization, and embodied adaptation. Retrieved operational examples serve as contextual references for action generation and are jointly conditioned on the target robot's embodied constraints to enable adaptive action execution. The approach ensures generated actions satisfy the physical and functional requirements of diverse robotic platforms, offering a coherent methodology for cross-embodied robotic operation [9].

3. Methodology

3.1. Overall Architecture

This chapter introduces a retrieval-enhanced vision-language-action (VLA) strategy for cross-haptic robot operations, designed to mitigate challenges in transferring operation strategies across robots with differing mechanical structures and control interfaces. Rather than relying exclusively on real-time visual inputs and language instructions to produce actions, the framework leverages retrieved historical operation experiences specific to the target robot, thereby enhancing cross-robot strategy adaptability [10]. As illustrated in Figure 1, the system comprises five core modules: multimodal state encoding, cross-haptic data synchronization, haptic perception retrieval, compatibility-gated aggregation, and target haptic action decoding.

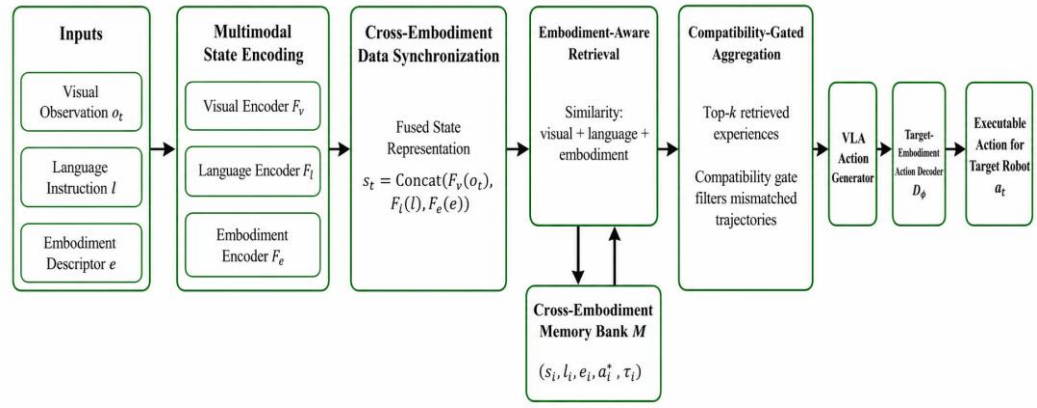

Figure 1. Architecture of the Proposed Enhanced VLA Retrieval Strategy

The visual encoder extracts object- and scene-level features from RGB or RGB-D observations [11]. The language encoder transforms task instructions into structured representations. The embodied encoder captures robot-specific attributes—including degrees of freedom, gripper configuration, action dimensionality, control frequency, and workspace boundaries. These encoded features are fused and jointly utilized for both historical experience retrieval and action generation.

3.2. Problem Definition

At time step t, the robot receives the current visual observation o_t, the language instruction l, and the robot's embodied description e. The model is required to predict the actions executable by the target robot given these inputs:

$$a_t = \pi_\theta(o_t, l, e, \mathcal{R}_t) \quad (1)$$

Here, π_θ denotes the vision-language-action strategy parameterized by θ, while $\mathcal{R}_t$ represents retrieved historical operation experience. In contrast to conventional vision-language-action strategies that rely solely on current observations and language instructions, this formulation incorporates retrieved operational experience as auxiliary information to enhance action generation.

3.3. Multimodal Encoding and Data Synchronization

The cross-haptic robot dataset typically exhibits variations in camera sampling frequency, control frequency, action dimensionality, and trajectory duration. To facilitate consistent comparison across robotic platforms, operational data must be temporally aligned and transformed into a standardized state-action representation. Specifically, visual features v_t, language features q, and haptic features b are extracted using the visual encoder F_v, language encoder F_l, and haptic encoder F_e, respectively, then fused to yield the current state representation:

$$s_t = \text{Concat}(F_v(o_t), F_l(l), F_e(e)) \quad (2)$$

Here, s_t denotes the fused state feature, and $\text{Concat}(\cdot)$ denotes feature concatenation. During training, expert-provided action labels a_t^* serve as supervision signals [12]. These labels may correspond to joint-level control commands, end-effector displacement vectors, or normalized action tokens. To accommodate differing robot

control frequencies, each action is temporally aligned to the nearest valid timestamp, and full trajectories are segmented into fixed-length windows.

3.4. Cross-Embodiment Memory Bank

The experience repository stores synchronized operational experiences acquired from diverse robot platforms.

$$\mathcal{M} = \{(s_i, l_i, e_i, a_i^*, \tau_i)\}_{i=1}^{N} \quad (3)$$

Here, $\mathcal{M}$ denotes the experience database, N the number of samples, s_i the state representation, l_i the task instruction, e_i the robot's embodied description, a_i^* the expert-provided action sequence, and τ_i the trajectory identifier. This structured organization enables retrieval of historical experiences based jointly on task relevance and structural compatibility across robot embodiments.

3.5. Embodiment-Aware Retrieval

Given the current query state s_t, the retrieval module computes the similarity between it and each sample in the experience database.

$$S_i = \alpha \cdot \mathrm{sim}(v_t, v_i) + \beta \cdot \mathrm{sim}(q, q_i) + \gamma \cdot \mathrm{sim}(b, b_i) \quad (4)$$

Here, S_i denotes the retrieval score; v_t and v_i represent the visual features of the current and candidate samples, respectively; q and q_i denote their language features; b and b_i encode their embodied features; and $\mathrm{sim}(\cdot)$ signifies cosine similarity. Unlike conventional semantic-only retrieval approaches, this module incorporates the robot's embodiment compatibility as a core retrieval criterion, thereby filtering out historical trajectories that match the task description linguistically but are physically infeasible for the target robot [13].

3.6. Compatibility-Gated Aggregation

Following top-k retrieval, the compatibility gating module dynamically modulates the influence of each retrieved sample on action generation.

$$g_i = \sigma(W_g[s_t; s_i; b; b_i] + c_g) \quad (5)$$

Here, g_i denotes the gating weight, $\sigma(\cdot)$ the Sigmoid function, W_g a learnable weight matrix, c_g a bias term, and $[s_t; s_i; b; b_i]$ the concatenated feature representation [14]. The gated fusion yields the refined retrieval representation:

$$r_t = \sum_{i=1}^{k} \frac{g_i \exp(S_i)}{\sum_{j=1}^{k} g_j \exp(S_j)} h_i \quad (6)$$

In this formulation, r_t denotes the final retrieval feature representation, h_i the encoded features of the i-th retrieved sample, and k the total number of retrieved samples. This mechanism attenuates the contribution of semantically similar but contextually mismatched experiences, thereby enhancing decision robustness by aligning sample relevance with the target robot's operational constraints.

3.7. Action Decoding and Optimization Objective

The VLA action generator fuses the current state s_t with the retrieved representation r_t to produce a general action representation z_t. Subsequently, an action decoder—parameterized by embodied knowledge of the target robot—maps this representation into the robot's executable action space:

$$\hat{a}_t = D_\phi(z_t, e) \quad (7)$$

Here, $\hat{a}_t$ denotes the predicted action, and D_ϕ denotes the action decoder parameterized by ϕ. Training employs joint optimization of multiple objectives:

$$\mathcal{L} = \lambda_1 \| \hat{a}_t - a_t^* \|_2^2 + \lambda_2 \mathcal{L}_{\mathrm{ret}} + \lambda_3 \mathcal{L}_{\mathrm{emb}} \quad (8)$$

In this formulation, $\mathcal{L}$ represents the total loss function; λ_1, λ_2, and λ_3 are weighting coefficients for distinct loss components. $\mathcal{L}_{\mathrm{ret}}$ optimizes retrieval ranking performance, while $\mathcal{L}_{\mathrm{emb}}$ regularizes action generation fidelity. During training, parameters theta and phi are updated by minimizing $\mathcal{L}$. At inference time, generated actions are adapted to the robot's operational specifications before being dispatched to the controller.

3.8. Algorithm Procedure

Algorithm 1. Retrieval-Augmented VLA Policy for Cross-Embodiment Manipulation

Input: training demonstrations $\mathcal{D}$, memory bank $\mathcal{M}$, observation o_t, instruction l, embodiment descriptor e, retrieval size k.

Output: executable target-robot action $\hat{a}_t$.

1: Initialize policy π_θ, decoder D_ϕ, and memory bank $\mathcal{M}$.
2: Synchronize visual frames, instructions, embodiment descriptors, and expert actions.
3: Encode o_t into visual feature v_t.
4: Encode l into language feature q.
5: Encode e into embodiment feature b.
6: Construct fused state representation s_t.
7: Compute retrieval score S_i for each memory item.
8: Select top- k retrieved examples.
9: Compute compatibility gate g_i for each retrieved item.
10: Aggregate retrieved features into r_t.
11: Fuse s_t and r_t in the VLA generator.
12: Generate general action representation z_t.
13: Decode z_t into target action $\hat{a}_t$.
14: Update θ and ϕ by minimizing $\mathcal{L}$ during training.
15: Output clipped executable action $\hat{a}_t$ during inference.

This method introduces three key innovations [15]. First, the retrieval process does not rely exclusively on visual or linguistic similarity when selecting historical experiences. Second, the compatibility gating mechanism mitigates interference from incompatible trajectories during action generation. Third, the action decoder tailors the generated outputs to the target robot's specific characteristics, thereby enhancing cross-robot deployment capability.

4. Results and Analysis

4.1. Experimental Setup

This experiment was evaluated using the Open X-Embodiment dataset [5, 9]. The primary objective was to assess how the embodied perception retrieval mechanism enhances cross-platform generalization capability in robotic manipulation tasks. This dataset comprises manipulation trajectory data collected from diverse robot platforms, fixture configurations, control modalities, and task environments, making it well-suited for investigating experience transfer across heterogeneous robotic systems. It is a publicly available large-scale robotic manipulation dataset, and its resources are distributed under the Apache-2.0 and CC-BY 4.0 license agreements.

To reduce computational overhead, a representative subset of tasks was selected for evaluation, including pick-and-place, object relocation, pushing, pulling, and basic tool-use operations. Robots with sufficient trajectory coverage were designated as source platforms, while others served as target platforms for generalization testing. The experience database included only source-platform data to prevent contamination of target test data. During baseline fine-tuning for target platforms, only a limited set of optimized target-platform trajectories was used; crucially, the final test trajectories were excluded from both model training and retrieval processes.

All experiments were conducted on a workstation equipped with an Intel i7-class CPU, 32 GB RAM, and an NVIDIA RTX-series GPU. Implementation was carried out in Python 3.10 with PyTorch [1, 9]. Training employed a batch size of 32, a learning rate of 1×10^(-4), and the AdamW optimizer. For the retrieval module, the top five most relevant samples were retrieved per query to provide auxiliary context, and loss function weighting parameters were set to λ_1 = 1.0, λ_2 = 0.3, and λ_3 = 0.2. Each experimental

configuration was independently repeated five times (n = 5). Statistical significance was assessed via paired t-tests, with $p < 0.05$ indicating a significant difference.

Data preprocessing included parsing RGB image frames, task instructions, robot state information, and action sequences; resizing input images to a consistent resolution; and normalizing action values to the interval [−1, 1]. Trajectories with incomplete or corrupted segments were excluded. Continuous trajectories were segmented into fixed-length windows aligned with control timestamps. The dataset was partitioned into training, validation, and test sets using trajectory-level splitting, with proportions of 70%, 15%, and 15%, respectively.

Four comparative methods were evaluated: BC-VLA, VLA augmented with language-only retrieval, VLA augmented with visual-language retrieval, and VLA followed by target-platform fine-tuning. Performance was quantified using three metrics: action prediction error (APE), cross-platform generalization gap (CEG), and retrieval compatibility score (RCS). APE measures prediction accuracy—lower values indicate higher fidelity; CEG quantifies performance degradation when transferring between platforms—smaller values reflect stronger generalization; RCS evaluates alignment between retrieved experiences and the target platform's operational context—higher scores denote greater relevance and compatibility.

4.2. Performance Comparison

Table 1 presents the comparison results of the proposed method with four baseline methods. The proposed method achieved the best performance across all evaluation indicators, though the magnitude of improvement varied by metric. A variant incorporating vision-language alignment and target-specific fine-tuning attained superior performance in the action prediction error (APE) metric, attributable to direct use of target-robot samples during training. However, its robot compatibility score (RCS) remained lower than that of the proposed method, suggesting that while target-robot fine-tuning enhances prediction accuracy in specific operational contexts, it exhibits limitations in effectively selecting and leveraging relevant historical experience.

Table 1. Performance Comparison Across Baselines, Mean ± SD, N=5

Method	APE ↓	CEG ↓	RCS ↑
BC-VLA	0.183 ± 0.017	0.131 ± 0.021	0.506 ± 0.036
VLA + Language Retrieval	0.167 ± 0.015	0.116 ± 0.018	0.582 ± 0.031
VLA + Vision-Language Retrieval	0.154 ± 0.012	0.097 ± 0.016	0.649 ± 0.029
VLA + Target Fine-tuning	0.143 ± 0.014	0.091 ± 0.015	0.614 ± 0.038
Proposed Method	0.134 ± 0.010	0.074 ± 0.013	0.709 ± 0.027

Relative to the strongest baseline, the proposed method reduced both action prediction error and cross-body generalization gap, demonstrating enhanced accuracy in action prediction and improved transferability across distinct robotic platforms. Statistical analysis confirmed that these improvements in APE and cross-body generalization gap (CEG) were statistically significant [5, 10]. Additionally, the method improved the retrieval compatibility score, indicating that integrating embodied robot information facilitates more effective filtering of historical experiences aligned with the target robot's physical and functional characteristics.

4.3. Ablation Study and Mechanism Verification

Table 2 presents the results of the ablation experiments. Removing the action adaptation decoder led to the largest increase in APE. Removing the embodied perception retrieval module caused a significant decrease in RCS. Removing the compatibility gating module resulted in an increase in CEG. Overall, these three modules serve distinct functions: the retrieval module enhances the matching quality of historical experiences,

the gating module improves the strategy's robustness against incompatible samples, and the action decoding module increases the executability of generated actions on the target robot.

Table 2. Ablation Study of the Proposed Method, Mean ± SD, N=5

Variant	APE ↓	CEG ↓	RCS ↑
Full Proposed Method	0.134 ± 0.010	0.074 ± 0.013	0.709 ± 0.027
w/o Embodiment-Aware Retrieval	0.151 ± 0.013	0.095 ± 0.017	0.638 ± 0.033
w/o Compatibility Gate	0.145 ± 0.012	0.087 ± 0.016	0.662 ± 0.030
w/o Action Adaptation Decoder	0.161 ± 0.016	0.108 ± 0.020	0.695 ± 0.026
w/o Retrieval Module	0.179 ± 0.018	0.128 ± 0.022	0.519 ± 0.039

The ablation results align with the methodological design outlined in Chapter 3. The superiority of the complete model does not stem from merely increasing input information, but from integrating the robot's embodied conditions to screen and leverage historical experiences better suited to the current task. Although the variant without the action decoder achieves stronger retrieval performance, it still fails to guarantee that retrieved trajectories can be directly executed on the target robot—demonstrating that both experience retrieval and action adaptation are essential components.

4.4. Convergence Analysis

Figure 2 illustrates the evolution of validation loss across 50 training cycles. Overall, the proposed method exhibits faster convergence than the BC-VLA baseline. During the initial phase—particularly within the first 10 cycles—fluctuations in the loss curve arise as retrieval results and gating weights undergo preliminary adjustment. As training proceeds, the loss steadily decreases, and the proposed method outperforms all baseline models after 20 cycles. By approximately cycle 30, the loss curve stabilizes, suggesting that a robust alignment has been established between the retrieval representation and the action decoding module.

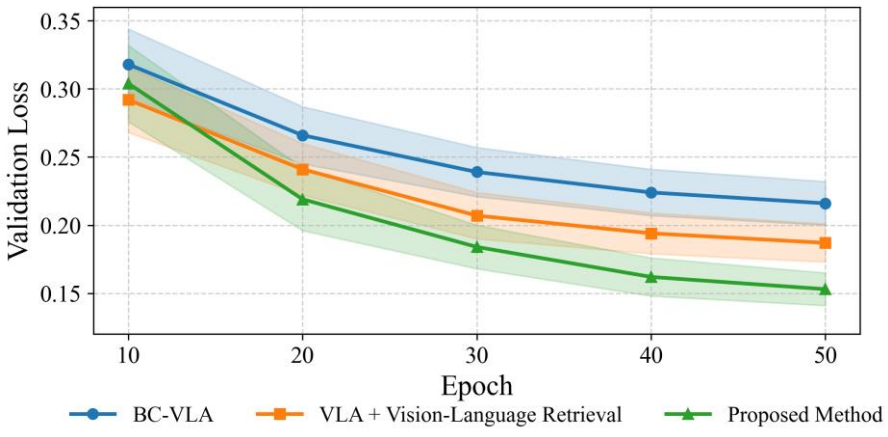

Figure 2. Validation Loss Convergence Curve with 95% Confidence Interval, N=5

Integration analysis reveals that the retrieval mechanism requires effective filtering to operate optimally. Although the visual language retrieval approach yields improvements over BC-VLA, its final validation loss remains higher than that of the proposed method. This discrepancy stems from residual retrieval of operation trajectories misaligned with the target robot's structural constraints. At cycle 50, the proposed method achieves a validation loss of 0.153 ± 0.012, compared to 0.187 ± 0.014 for VLA combined

with visual language retrieval and 0.216 ± 0.016 for BC-VLA—representing reductions of 18.2% and 29.2%, respectively. The compatibility gating mechanism mitigates such mismatches by attenuating the contribution of irrelevant examples during action generation.

4.5. Stability, Robustness, and Significance Analysis

Figure 3 shows the average percentage error (APE) under five excluded target robot settings. Overall, the proposed method achieves lower average error and reduced result fluctuation compared to the baseline method; however, performance varies across target robots. Notably, all methods exhibit relatively high APE on target 4, likely attributable to differences in fixture configuration and camera viewpoint associated with this robot. This suggests that embodied-information-based retrieval enhances model adaptability across robots, yet remains constrained when substantial physical disparities exist among platforms.

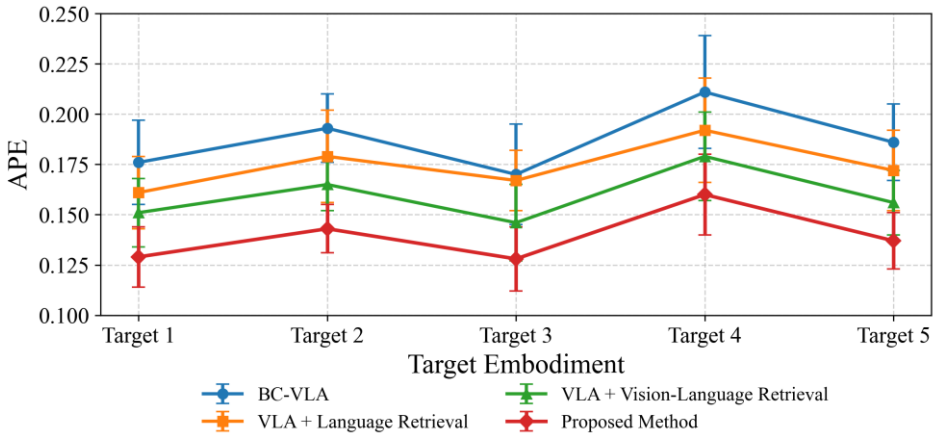

Figure 3. APE Across Target Embodiments with 95% Confidence Interval, N=5

Across varying target robot configurations, the proposed method consistently reduces mean APE while exhibiting narrower result variability. Paired statistical comparison against the VLA plus visual-language retrieval approach yields significance ($p = 0.027$), confirming not only superior overall accuracy but also improved stability during cross-robot transfer.

The RCS index highlights the method's efficacy in experience selection: the model preferentially retrieves historical trajectories relevant to the current task and better aligned with the target robot's embodiment, thereby grounding action generation in contextually appropriate prior experiences. From a robustness perspective, the compatibility gating mechanism mitigates interference from noisy or mismatched trajectory samples. Regarding data transparency, experiments rely exclusively on publicly available datasets, accompanied by explicit licensing documentation, trajectory-level dataset partitioning, and fully reproducible preprocessing protocols. The elevated errors observed for Target 4 further indicate that method performance remains sensitive to the representational breadth of the experience library—significant structural divergence between target and source robots continues to challenge effective experience matching and action adaptation [12].

5. Conclusion

This study investigated whether the enhanced retrieval VLA strategy could enable robots to transfer operational experiences learned on one robot to others. Experimental results demonstrated that incorporating the target robot's own structural and behavioral characteristics significantly improved the selection and utilization of relevant historical experiences, thereby enhancing cross-robot transfer performance.

The experiments revealed that relying solely on task descriptions or image similarity for case retrieval does not guarantee applicability to the current robot. Even when performing identical tasks, robots with differing mechanical configurations, end-effector designs, or control interfaces often require distinct action sequences. Consequently, effective experience retrieval must jointly consider task semantics and robot-specific physical and operational attributes.

When filtered historical operation experiences were integrated into action generation, model performance improved across diverse robot platforms. Relative to the baseline strategy, the proposed approach achieved consistently lower action prediction errors and greater behavioral stability under varied target robot configurations. These outcomes indicate that judicious reuse of prior operational knowledge supports robust adaptation to inter-robot heterogeneity and strengthens strategy transfer efficacy.

Ablation studies confirmed the critical role of the action adaptation decoder in cross-robot execution. Due to disparities in kinematic structure, actuation range, and control granularity, raw action outputs from source robots cannot be directly applied to target platforms. Removing this module led to a marked decline in performance, underscoring that action-space alignment—tailored to the target robot's capabilities—is essential for executable and reliable policy transfer.

Several limitations remain. The evaluation relied exclusively on offline operation data from a public robot dataset; real-world deployment introduces dynamic environmental changes and long-term operational drift. The range of robot morphologies tested was constrained and does not encompass the full spectrum of industrial or research platforms. Additionally, the retrieval mechanism incurs computational overhead during inference, and its effectiveness is contingent on both the volume and representational coverage of experiences stored in the database.

Future work should validate the method in live robotic systems and develop mechanisms for continuous experience accumulation and refinement during operation. For robots with substantial structural divergence, the experience filtering strategy requires refinement to better align retrieved cases with target-platform constraints. Integrating complementary signals—such as human feedback or real-time error diagnostics—may further improve adaptability in complex or failure-prone scenarios.

References

1. Y. Zhang, R. Wang, J. Lin, Z. Wang, and X. Qi, "Retrieval-VLA: Training-Free In-Context Adaptation for Vision-Language-Action Models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2026, pp. 1358–1367.
2. G. Sarch, Y. Wu, M. J. Tarr, and K. Fragkiadaki, "Open-ended instructable embodied agents with memory-augmented large language models," arXiv preprint arXiv:2310.15127, 2023.
3. D. Wu, X. Wei, G. Chen, H. Shen, X. Wang, W. Li, and B. Jin, "Generative multi-agent collaboration in embodied AI: A systematic review," arXiv preprint arXiv:2502.11518, 2025.
4. Z. Xie, J. Gong, X. Tan, Z. Zhang, and Y. Xie, "Zero-shot robotic manipulation via 3D Gaussian splatting-enhanced multimodal retrieval-augmented generation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 40, no. 22, Mar. 2026, pp. 18683–18691.
5. T. Tsuji, Y. Kato, G. Solak, H. Zhang, T. Petrič, F. Nori, and A. Ajoudani, "A survey on imitation learning for contact-rich tasks in robotics," *Int. J. Robot. Res.*, 2025, p. 02783649261417694.
6. J. Zhou, K. Ye, J. Liu, T. Ma, Z. Wang, R. Qiu, et al., "Exploring the limits of vision-language-action manipulation in cross-task generalization," in *Adv. Neural Inf. Process. Syst.*, vol. 38, 2026, pp. 139899–139927.
7. Z. Yin, X. Wang, J. Jiang, K. Deng, P. Chen, S. Li, et al., "MiVLA: Towards Generalizable Vision-Language-Action Model with Human-Robot Mutual Imitation Pre-training," arXiv preprint arXiv:2512.15411, 2025.
8. O. Farid and H. A. Mahmoud, "Vision Language Action Models for Embodied Intelligence: A Structured Taxonomy, Critical Analysis and Future Research Directions," *Comput. Discovery Intell. Syst.*, vol. 3, no. 2, pp. 91–131, 2026.
9. X. Li, P. Li, L. Qian, M. Liu, D. Wang, J. Liu, et al., "What matters in building vision–language–action models for generalist robots," *Nat. Mach. Intell.*, vol. 8, no. 2, pp. 158–172, 2026.
10. Y. Zhang, Z. Gao, S. Li, L. H. Chen, K. Liu, R. Cheng, et al., "RoboWheel: A Data Engine from Real-World Human Demonstrations for Cross-Embodiment Robotic Learning," arXiv preprint arXiv:2512.02729, 2025.
11. W. Weng, T. Wu, L. Chen, S. Xie, Z. Wang, X. Xu, et al., "Language-Grounded Decoupled Action Representation for Robotic Manipulation," arXiv preprint arXiv:2603.12967, 2026.

12. H. Song, L. Wang, X. Qiao, Y. Chen, D. Sun, and Z. Sun, "Embodied intelligence for robot manipulation: development and challenges," *Vicinagearth*, vol. 2, no. 1, p. 8, 2025.
13. J. Liu, H. Chen, P. An, Z. Liu, R. Zhang, C. Gu, et al., "HybridVLA: Collaborative diffusion and autoregression in a unified vision-language-action model," arXiv preprint arXiv:2503.10631, 2025.
14. Open X-Embodiment Collaboration, A. O'Neill, A. Rehman, A. Gupta, A. Maddukuri, A. Gupta, et al., "Open X-Embodiment: Robotic learning datasets and RT-X models," arXiv preprint arXiv:2310.08864, vol. 1, no. 2, 2023.
15. J. Wen, Y. Zhu, J. Li, M. Zhu, Z. Tang, K. Wu, et al., "TinyVLA: Towards fast, data-efficient vision-language-action models for robotic manipulation," *IEEE Robot. Autom. Lett.*, 2025.